Transactions of the Korean Nuclear Society Spring Meeting
Jeju, Korea, May 23-24, 2019

Considerations on Machine Learning Programming of Multiple Linear Regression

Yong Suk Suh*, Seung Ki Shin, Dane Baang, Jong Bok Lee
Research Reactor System Design Div., Korea Atomic Energy Research Institute (KAERI), 111, Daedeok-Daero 989Beon-Gil, Yuseong-Gu, Daejeon, 34057, Korea
**Corresponding author: yssuh@kaeri.re.kr*

1. Introduction

A machine learning programming (MLP) of simple linear regression (SLR) was briefly reviewed in a previous paper [1]. The paper showed that the outcome of the programming depends on the initial value, learning rate and epoch (the number of trainings) that are determined by its programmer. The SLR can be limited or biased to the variable because it uses only one independent variable. To overcome this limitation, we can use the multiple linear regression (MLR) to predict the value of dependent variable given multiple independent variables.

The MLR can be applied to the nuclear facility if the dependent and independent variables are properly selected. For example, we can select several process parameters as the independent variables and predict a phenomenon as the dependent using the MLR. Prior to applying the MLR to the nuclear facility, it is require to analyze the domain to be predicted and variables to be regressed. However, this paper selects arbitrary values for the independent and dependent. The purpose of this paper is to review characteristics of the MLP of MLR and investigate what should be considered for the MLP of MLR.

2. Multiple Linear Regression (MLR)

The MLR is used for reducing the bias of SLR and is simply represented as follows:

$$y = w_1 x_1 + w_1 x_1 + ... + w_n x_n + b, \quad (1)$$

where y is a dependent variable, x is an independent variable, and w and b are parameters to be calculated using the regression: w is a slope of y and b is a y-intercept.

The MLP in this paper uses Tensorflow libraries under an Anaconda development environment that is identical to [1]. The regression uses the mean square error method. The parameter optimization uses the gradient descent method.

For MLP of MLR in this paper, the arbitral values 1, 2, 3, 4 and 5 are chosen for the dependent variable, y. The arbitral values TC1, TC2 and TC3 in Table 1 are chosen for the independent variable, x. The TC1 is a test case for SLR, not for MLR. The TC1 is to compare with the TC2 to show that the MLR with the TC2 works well. The x values in TC1 is divided into two parts x_1 and x_2 in TC2 to test the MLR and verify whether we could obtain the same outcome as the TC1. For TC3, x_2 is generated by multiplying x_2 of TC2 by 1,000 in order to make a big difference in scale between the x_1 and x_2.

Table 1. Arbitrary test data for MLR

Test Case	Arbitrary x values as training data					
TC1	x	1	2	3	4	5
TC2	x_1	1	0	3	0	5
	x_2	0	2	0	4	0
TC3	x_1	1	0	3	0	5
	x_2	0	2000	0	4000	0

As shown in Table 2, The TC1 and TC2 generate almost same results. They are also the same results calculated by the MS-excel. However, The MLP cannot calculate the TC3 as the MS-excel does. The "nan" in TC3 of Table 2 denotes "not a number", indicating that the MLP cannot generate the result. When we make the big difference in scale between the x_1 and x_2, the MLP could not generate the results although we run (train) it with 1e-4 learning rate and 1e6 epochs. This is due to the big difference in scale between the x_1 and x_2.

Table 2. Results of TC1, TC2 and TC3

Test Case	w_1, w_2 (MS-excel)	b (MS-excel)	cost	learning rate	epoch
TC1	1 (1)	0 (0)	0	0.1	1
TC2	0.99, 0.99 (1, 1)	1.6e-7 (-8.8e-16)	1.4e-14	0.1	1e3
TC3	nan, nan (1, 0.001)	nan (-2.2e-16)	nan	1e-4	1e6

To obtain the result of TC3, we need to make the x_1 and x_2 the same scale [-1, 1] using the Equation (2) that normalize the x_1 and x_2 as shown in Table 3.

$$\text{Normalization}(x_i) = \frac{(x_i - \bar{x}_i)}{(x_{imax} - x_{imin})}, \quad (2)$$

where $\bar{x}_i$ is a mean of dependent variable x_i, and x_{imax} and x_{imin} are the maximum and minimum of x_i.

Table 3. Normalized x values of TC3

TC3	x_1	-0.16	-0.36	0.24	-0.36	0.64
	x_2	-0.3	0.2	-0.3	0.7	-0.3

Using the normalized x_1 and x_2 values, we obtain the values of parameters w_1, w_2 and b as shown in Table 4, which shows different cost values depending on the learning rate and epochs. The values of y also depend on the values of the parameters as shown in Table 5.

In Table 4, as MS-excel calculates them, we can correctly obtain 1, 2, 3, 4 and 5 for the dependent variable only in TC3N2, TC3N3 and TC3N7. When setting the learning rate to between 0.1 and 0.9, the minimum epoch is 1e3. If the learning rate is greater than 1, we cannot

obtain the correct result. When learning rate is less than 0.01, the epoch shall be greater than 1e6. In conclusion, the outcome of the MLP of MLR depends on the learning rate and epoch that are determined by its programmer.

Table 4. Results of TC3 using the normalized x values

Test Result	w_1, w_2 (MS-excel)	b (MS-excel)	cost	learning rate	epoch
TC3N1	3.1982, 2.2750(5, 4)	2.9999 (3)	0.2986	0.1	1e2
TC3N2	4.9995, 3.9996	2.9999	1.5e-8	0.1	1e3
TC3N3	4.9999, 3.9999	3.0	1.5e-13	0.9	1e3
TC3N4	4.9999, 3.9999	0.9999	4.0	1.0	1e3
TC3N5	4.9999, 3.9999	0.9999	4.0	1.0	1e6
TC3N6	3.4531, 2.5260	2.9999	0.29	0.01	1e3
TC3N7	4.9999, 3.9999	2.9999	3.4e-9	0.01	1e6

Table 5. The values of y per Table 4

TC3N1	1.805	2.303	3.085	3.441	4.364
TC3N2	1.000	2.000	3.000	3.999	4.999
TC3N3	1.000	2.000	3.000	3.999	4.999
TC3N4	-1.00	-6e-7	9e-1	1.999	2.999
TC3N5	-1.00	-8e-6	9e-1	1.999	2.999
TC3N6	1.689	2.262	3.070	3.525	4.452
TC3N7	1.000	2.000	2.999	3.999	4.999

When MLR is used for the prediction, the multi-collinearity between the independent variables must be checked. Multi-collinearity is defined as the linearity correlation between independent variables [2]. Variance inflation factor (VIF) is used to measure the degree of multi-collinearity using the Equation (3). When Multi-collinearity increases, Variance of MLR is inflated. This is what the VIF means.

$$VIF_j = \frac{1}{1-R_j^2}, \qquad (3)$$

where R_j^2 is a goodness-of-fit measure for linear regression models between x_j and the other independent variables excluding x_j. R_j^2 is calculated using the MLR by assigning x_j as a dependent variable and other x variables as independent variables. R_j^2 ranges from 0 to 1. If R_j^2 is 1, the regression line perfectly fits.

Therefore, when R^2 is closed to 1, the VIF is close to infinity, which means that the multi-collinearity is stronger. From the statistics viewpoint, this reduces the prediction capability of MLR so that the independent variables shall be adjusted to decrease the VIF or the high multi-collinearity variable can be eliminated from the model. It is not necessary to duplicate the independent variables in the model. When VIF is greater than 10, the independent variable should be eliminated. The VIF of TC2, TC3 and TC4 is 1.9, which is proper for MLR.

In terms of MLR, both MS-excel and MLP can calculate the parameter w and b although the VIF is very high. For example, when VIF of TC4 and TC5 is 121, which is bad, the parameters, w and b, are calculated as shown in Table 8. The TC5 is normalized for the MLP of MLR as shown in Table 7. The VIF of TC5N is 145.

Table 6. Arbitrary test data with high VIF for MLR

Test Case	Arbitrary x values as training data					
TC4	x_1	1	3	5	7	9
	x_2	0.1	0.3	0.5	0.7	1
TC5	x_1	1	3	5	7	9
	x_2	10	30	50	70	100

Table 7. Normalized x values of TC5

TC5N	x_1	-0.5	-0.25	0	0.25	0.5
	x_2	-0.47	-0.24	0.02	0.2	0.53

Table 8. Results of TC4, TC5 and TC5N

Test Case	w_1, w_2 (MS-excel)	b (MS-excel)	cost	learning rate	epoch
TC4	0.5017, -0.0157 (0.5, 0)	0.4995 (0.5)	1.9e-7	0.001	1e5
TC5	nan, nan (0.5, 0)	Nan (0.5)	nan	0.001	1e5
TC5N1	2.6001, 1.4260 (4,0)	2.9999 (3)	1.9e-3	0.001	1e5
TC5N2	3.9941, 0.0059	2.9999	3.4e-8	0.01	1e6

Although the multi-collinearity in the independent variables exists, the MLP of MLR correctly calculates the parameters. As the statics worried about the effect of multi-collinearity, it is necessary to consider the elimination of the unnecessary independent variables by calculating the VIF of each independent variables.

3. Conclusions

For machine learning programming of multiple linear regression, normalization should be considered for different scales of independent variables. The machine learning program also generates different results depending on the learning rate and epoch as the simple linear regression does. Although multi-collinearity in the independent variables exists, it does not affect the generation of outputs.

REFERENCES

[1] Yong Suk Suh, et al., A Brief Review of a Machine Learning Programming of Simple Linear Regression, Transactions of the Korean Nuclear Society Spring Meeting, Jeju, Korea, May 17-18, 2018.
[2] http://en.m.wikipedia.org/wiki/Multicollinearity